Jena University Language & Information Engineering Lab

MODELL ROMANTiK
Variation • Reichweite • Aktualität

Automatische Analyse diachroner Wortsemantik

Johannes Hellrich, M.A.

Digital Humanities in Jena 23.11.2017

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

DFG

Wortsemantik und Wortumfeld

„Die Bedeutung eines Wortes ist sein Gebrauch in der Sprache.“

Wittgenstein, Philosophische Untersuchungen, 1953

“You shall know a word by the company it keeps!”

Firth, A synopsis of Linguistic Theory, 1957

Wortsemantik und Wortumfeld

Er liest ein Gedicht.
Susanne liest einen Roman.
Der Roman hat 100 Seiten.
Ihr Gedicht hat 3 Seiten.
Susanne hört eine Oper.
Peter hört ein Lied.
Das Lied ist in d-Moll.
Die Oper ist in d-Moll.

Wortsemantik und Wortumfeld

Er liest ein Gedicht.
Susanne liest einen Roman.
Der Roman hat 100 Seiten.
Ihr Gedicht hat 3 Seiten.
Susanne hört eine Oper.
Peter hört ein Lied.
Das Lied ist in d-Moll.
Die Oper ist in d-Moll.

Wortsemantik und Wortumfeld

Er liest ein Gedicht.
Susanne liest einen Roman.
Der Roman hat 100 Seiten.
Ihr Gedicht hat 3 Seiten.
Susanne hört eine Oper.
Peter hört ein Lied.
Das Lied ist in d-Moll.
Die Oper ist in d-Moll.

Wortsemantik und Wortumfeld

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	…
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

Spezifisches Wortumfeld

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	...
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

$$PMI := log(\frac{P(w,c)}{P(w)P(c)})$$

Spezifisches Wortumfeld

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	...
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

$$PMI := log(\frac{P(w,c)}{P(w)P(c)})$$

Spezifisches Wortumfeld

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	...
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

$$PMI := log(\frac{P(w,c)}{P(w)P(c)})$$

Spezifisches Wortumfeld

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	…
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

$$PMI := log(\frac{P(w,c)}{P(w)P(c)})$$

“If A and B have some environments in common and some not (e.g. oculist and lawyer) we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments.”

Harris, Distributional Structure, 1954

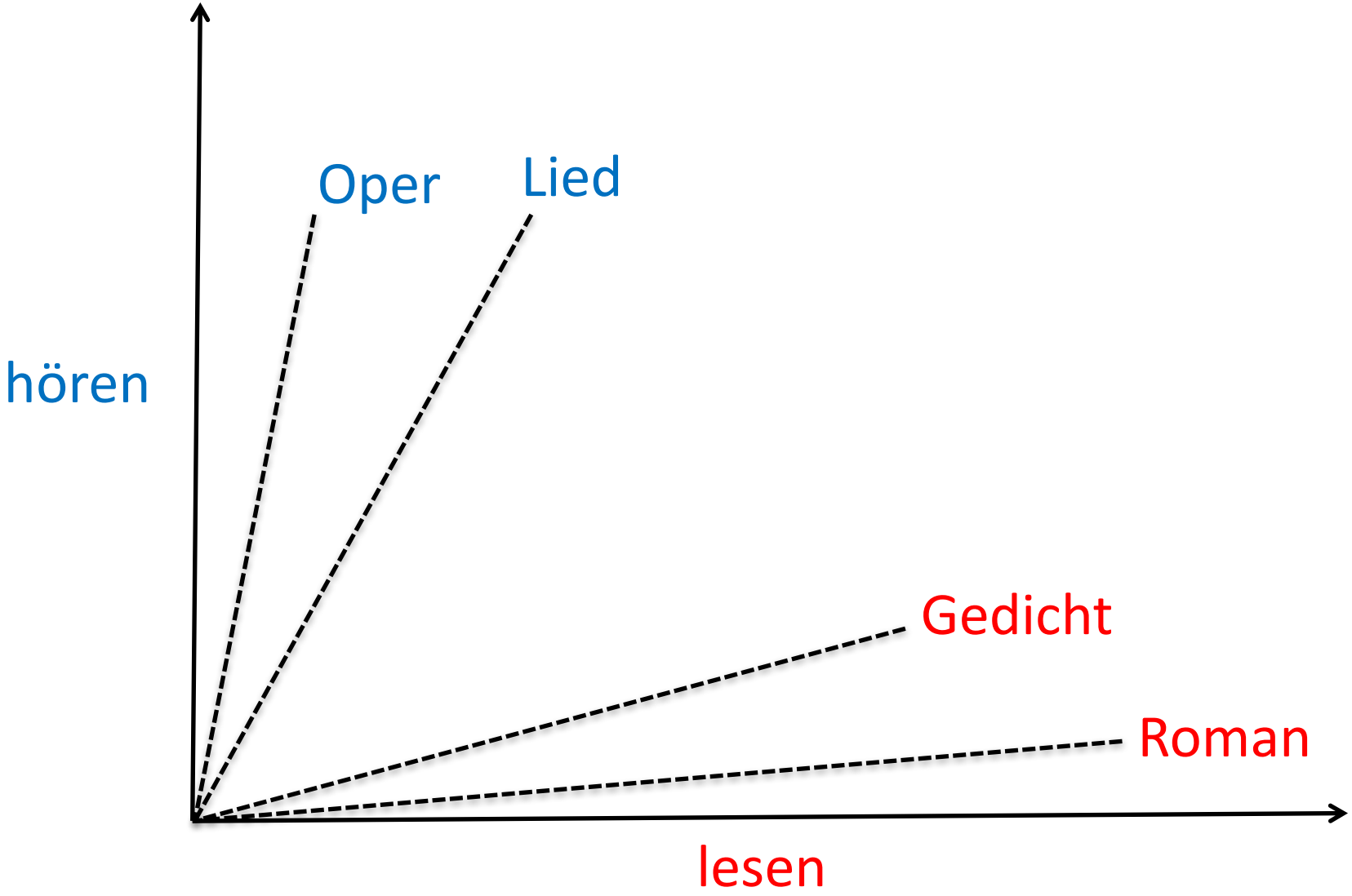
Wortähnlichkeit und Wortumfeld

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	…
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

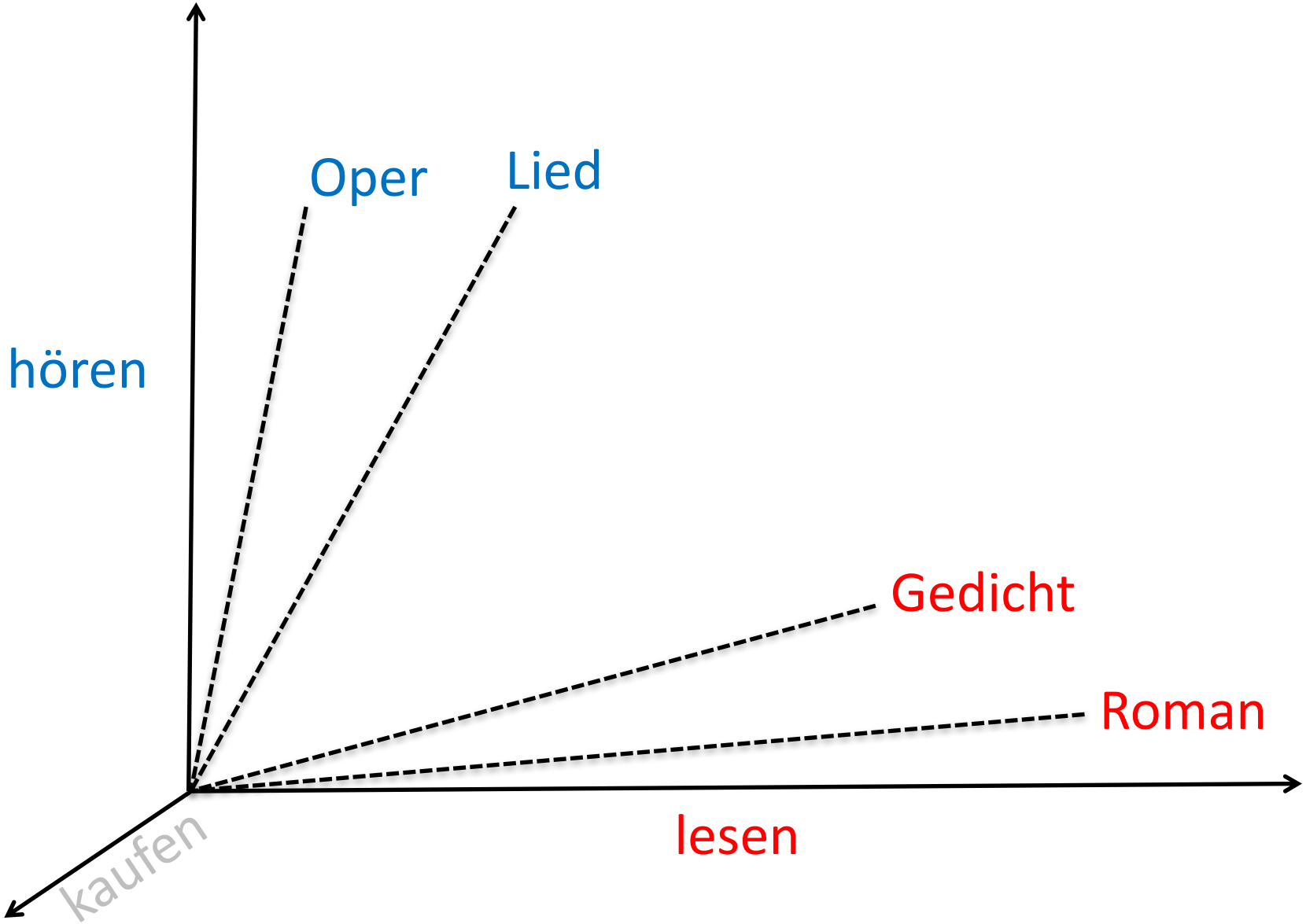
Wortähnlichkeit und Wortumfeld

	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	…
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

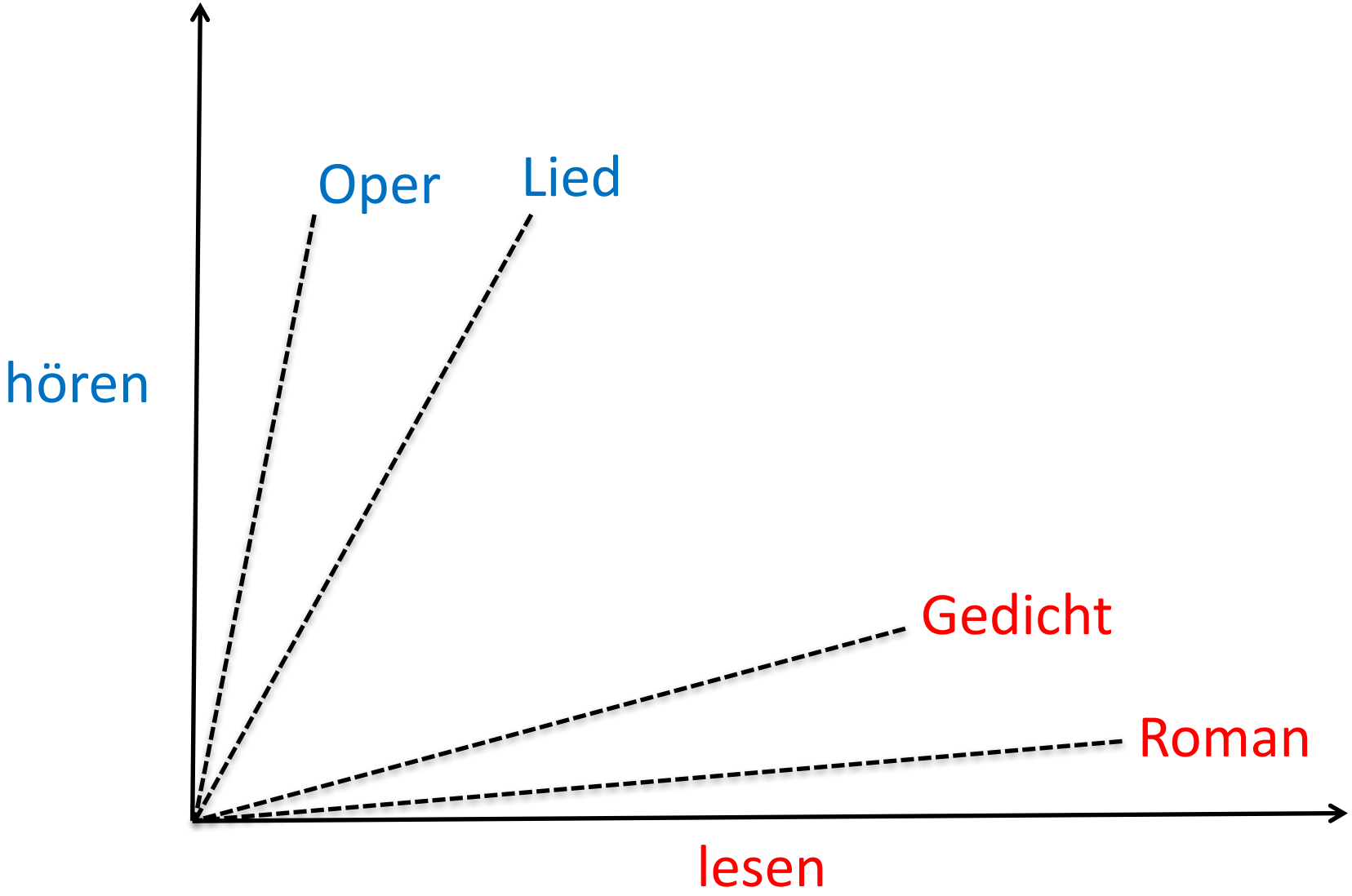
Wortähnlichkeit und Wortumfeld



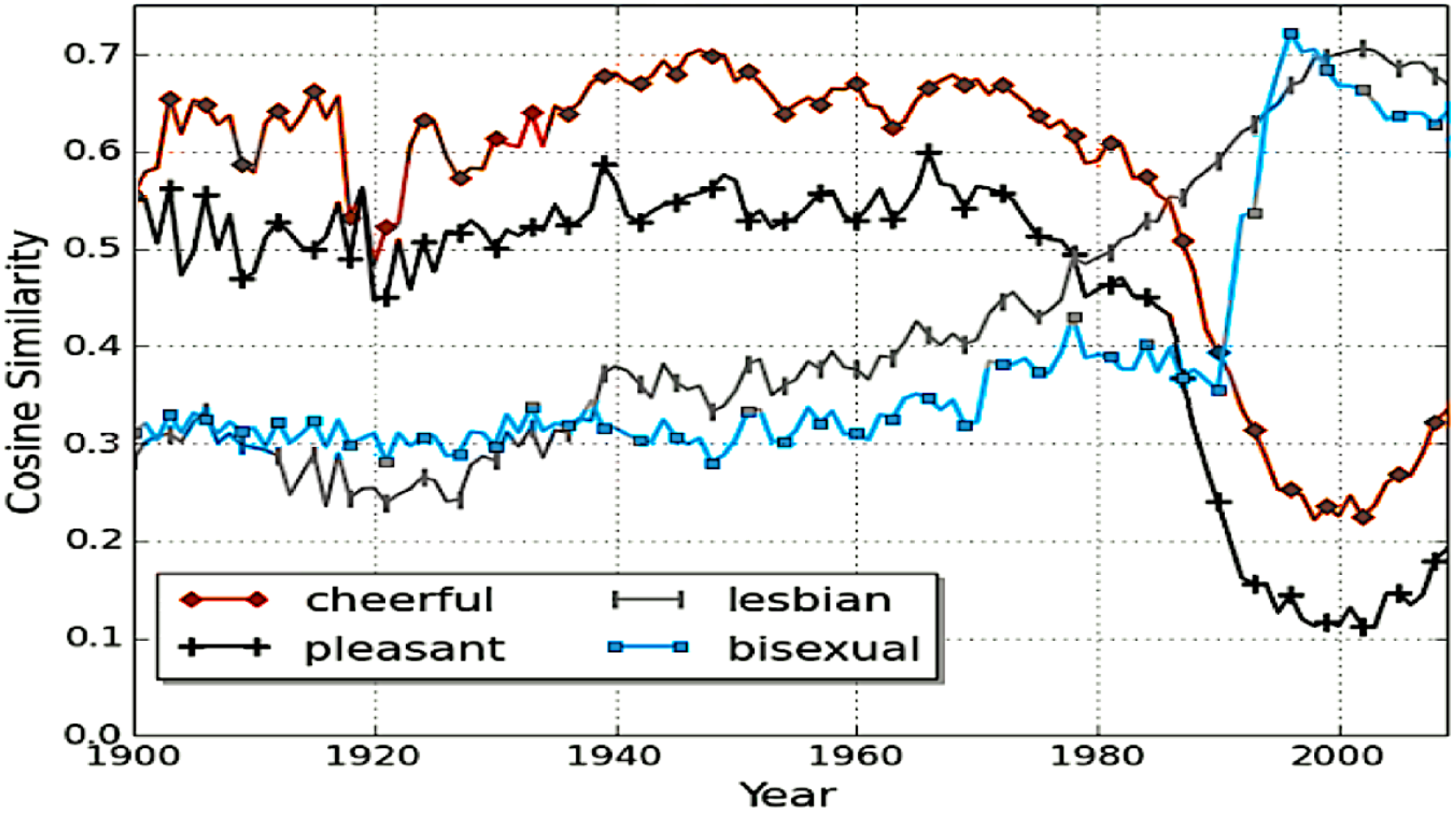
Wortähnlichkeit und Wortumfeld



Wortähnlichkeit und Wortumfeld

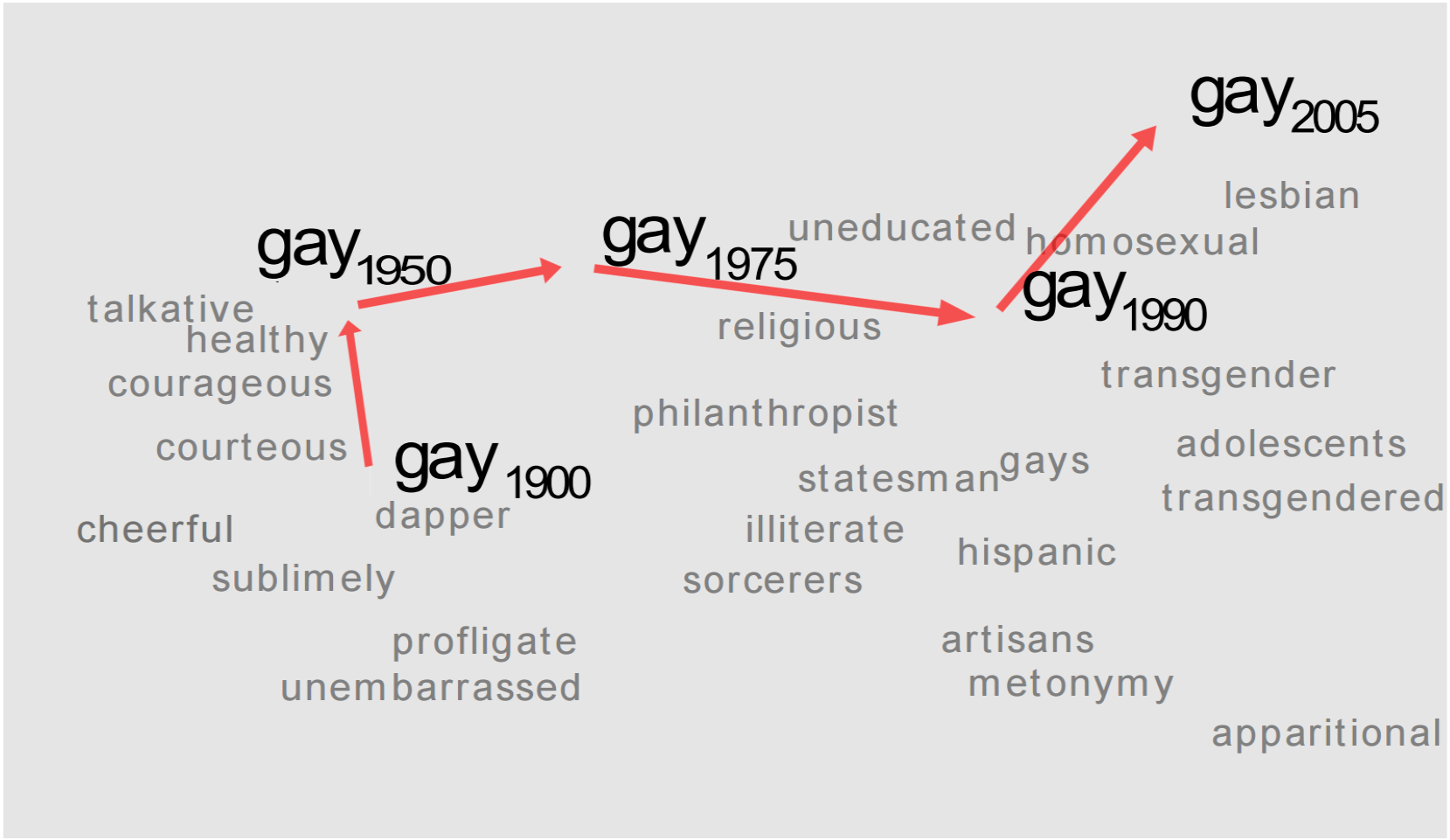


Messung diachroner Semantik: *gay* im 20. Jhdt.



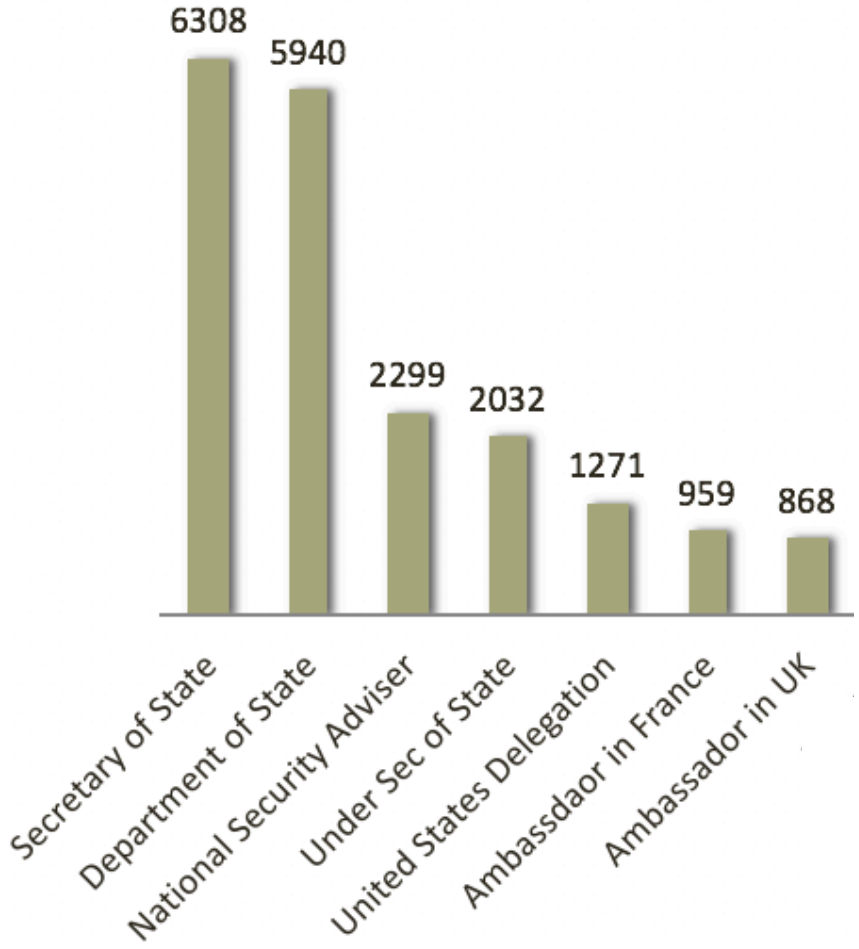
Kim et al. (2014): Temporal analysis of language through neural language models. In: *Proceedings of the Workshop on Language Technologies and Computational Social Science @ ACL 2014, 2014*, pp. 61–65

Ein Forschungstrend...



Kulkarni et al. (2015): Statistically significant detection of linguistic change. In: Proceedings of the 24th International Conference on World Wide Web: Technical Papers. pp. 625–635.

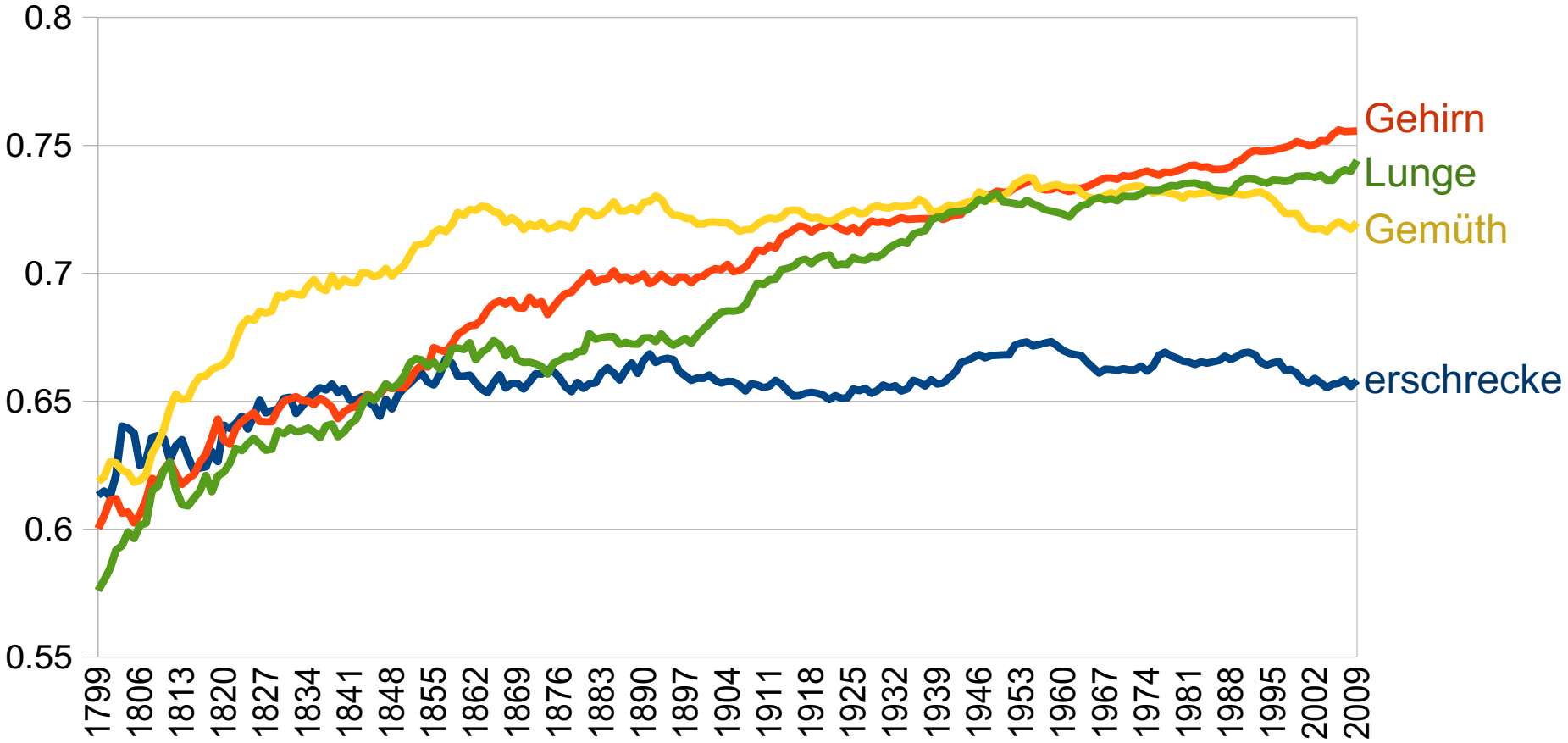
Ein Forschungstrend...



1860 - economy	1950 - economy
'management'	'economies'
'morality'	**'expanding'**
'self-government'	'domestic'
'utility'	**'balance'**
'administrative'	**'healthy'**
'education'	**'growth'**
'activity'	'industry'
'study'	**'stability'**
'economical'	**'expansion'**
'discipline'	'internal'

Jo (2016): Diplomatic history by data. Understanding Cold War foreign policy ideology using networks and NLP. In: Digital Humanities 2016. pp. 582–585.

Ein Forschungstrend...



Hellrich & Hahn (2016): Measuring the dynamics of lexico-semantic change since the German Romantic period. In: Digital Humanities. pp. 545–547.

… für Technikexperten

```
>>> model = Word2Vec(sentences, size=100, window=5, min_count=5, workers=4)
```

Persist a model to disk with:

```
>>> model.save(fname)
>>> model = Word2Vec.load(fname)  # you can continue training with the loaded model!
```

The word vectors are stored in a KeyedVectors instance in model.wv. This separates the read-only word vector lookup operations in KeyedVectors from the training code in Word2Vec:

```
>>> model.wv['computer']  # numpy vector of a word
array([-0.00449447, -0.00310097,  0.02421786, ...], dtype=float32)
```

The word vectors can also be instantiated from an existing file on disk in the word2vec C format as a KeyedVectors instance:

```
NOTE: It is impossible to continue training the vectors loaded from the C format because hidden weights, vocabulary frequenc

    >>> from gensim.models.keyedvectors import KeyedVectors
    >>> word_vectors = KeyedVectors.load_word2vec_format('/tmp/vectors.txt', binary=False)  # C text format
    >>> word_vectors = KeyedVectors.load_word2vec_format('/tmp/vectors.bin', binary=True)  # C binary format
```

You can perform various NLP word tasks with the model. Some of them are already built-in:

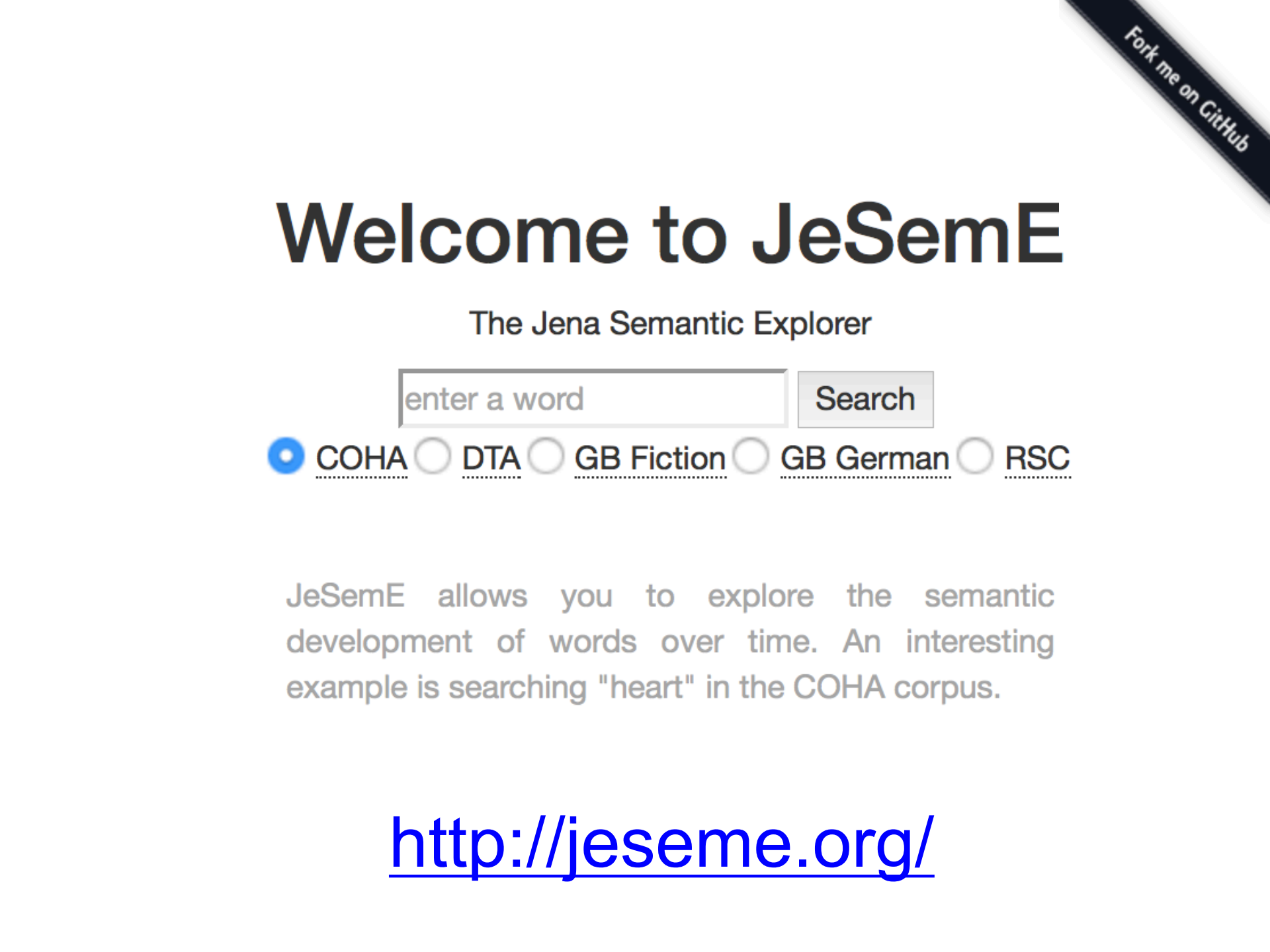
```
>>> model.wv.most_similar(positive=['woman', 'king'], negative=['man'])
[('queen', 0.50882536), ...]

>>> model.wv.most_similar_cosmul(positive=['woman', 'king'], negative=['man'])
[('queen', 0.71382287), ...]


>>> model.wv.doesnt_match("breakfast cereal dinner lunch".split())
'cereal'

>>> model.wv.similarity('woman', 'man')
0.73723527
```

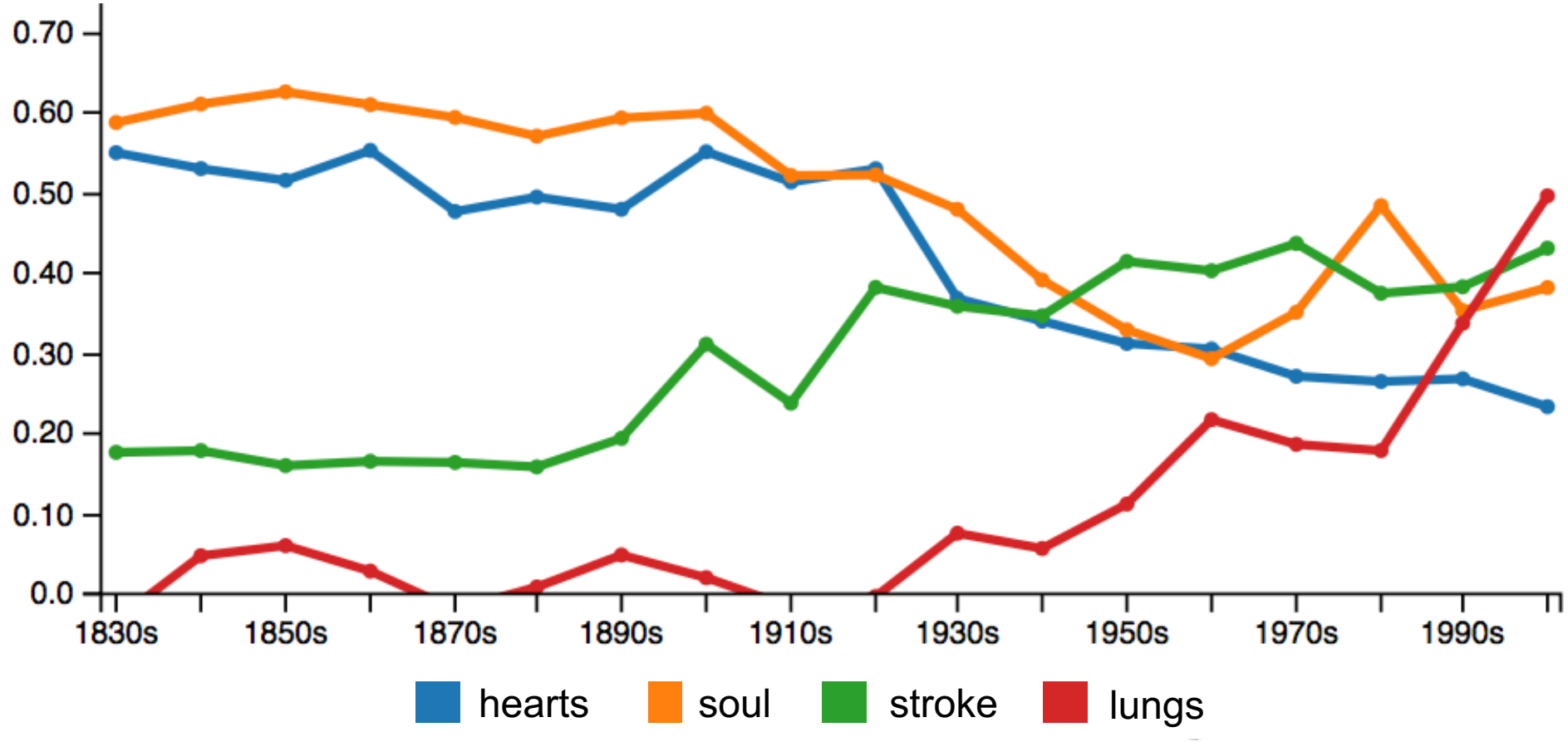
https://radimrehurek.com/gensim/models/word2vec.html


Fork me on GitHub
Welcome to JeSemE
The Jena Semantic Explorer
enter a word
Search
COHA
DTA
GB Fiction
GB German
RSC
JeSemE allows you to explore the semantic development of words over time. An interesting example is searching "heart" in the COHA corpus.
http://jeseme.org/

Korpora in JeSemE

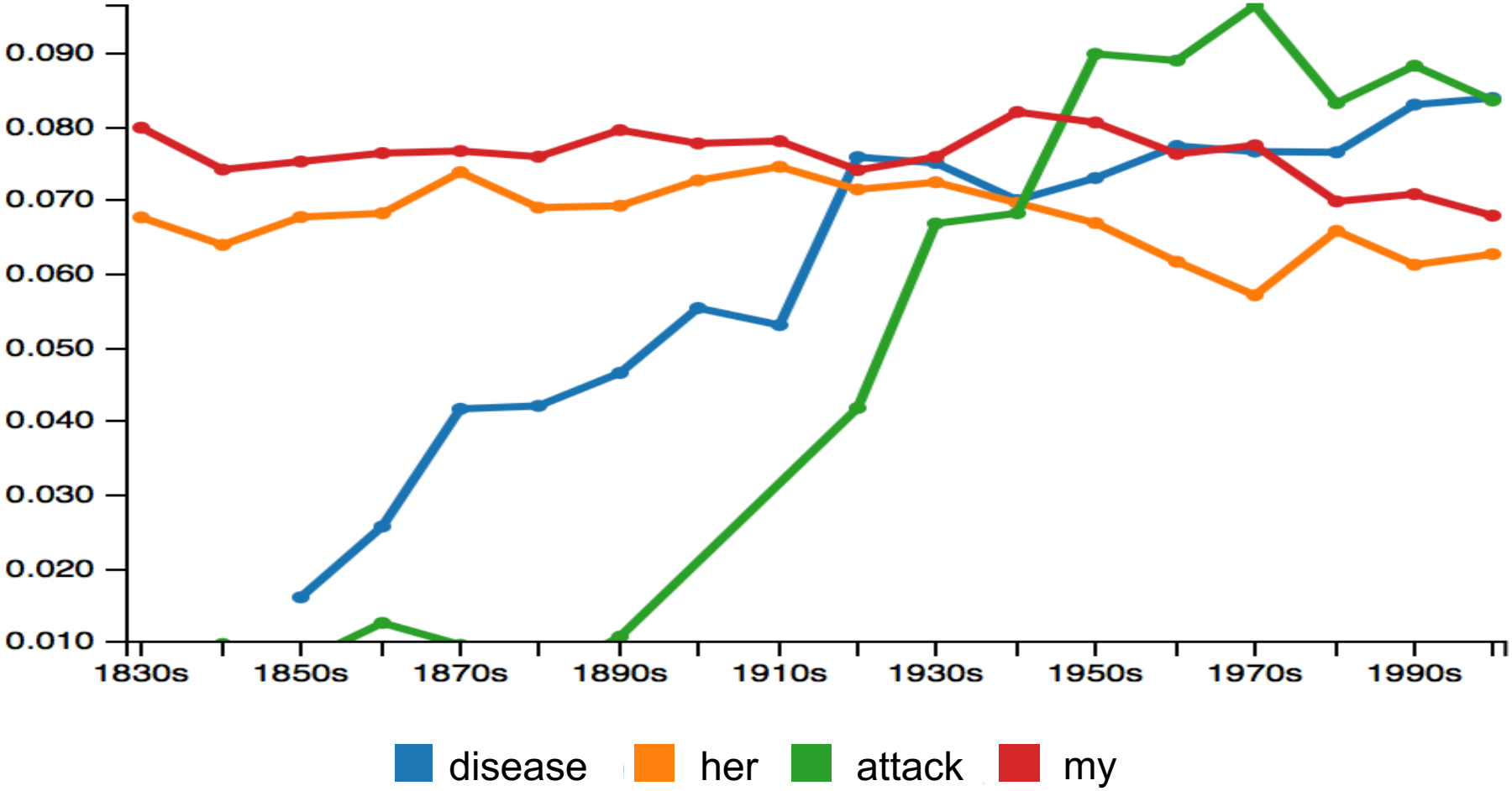
Korpus	Zeitraum	Wörter	Modellierte Wörter
Corpus of Historical American English	1830–2009	10^8	5.101
Deutsches Text Archiv	1751–1900	10^7	5.338
Google Book Fiction	1820–2009	10^{10}	6.492
Google Book German	1830–2009	10^9	4.449
Royal Society Corpus	1750–1869	10^7	3.080

Ähnlichste Wörter für *heart*



Johannes Hellrich & Udo Hahn: Exploring Diachronic Lexical Semantics with JeSemE. In: *ACL 2017, System Demonstrations.* pp. 31-36.

Spezifisches Umfeld für *heart*



Johannes Hellrich & Udo Hahn: Exploring Diachronic Lexical Semantics with JeSemE. In: *ACL 2017, System Demonstrations.* pp. 31-36.

Zu viele Wortumfelder

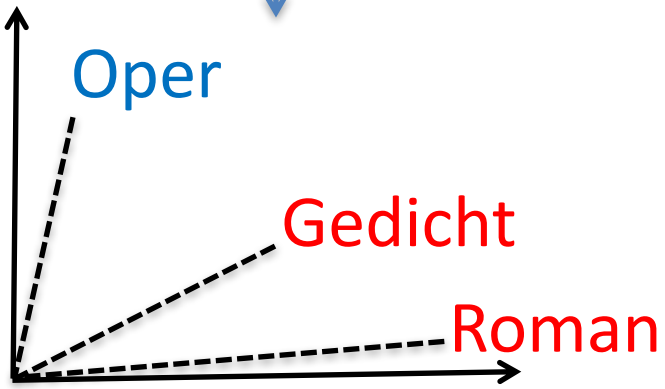
	lesen	Seiten	kaufen	essen	hören	
Roman	98	60	3	0	2	
Gedicht	67	10	1	0	8	…
Oper	4	8	0	0	38	
Lied	12	1	2	0	47	
⋮						

50.000 x 50.000 Wörter
= 2,5 Milliarden Kombinationen

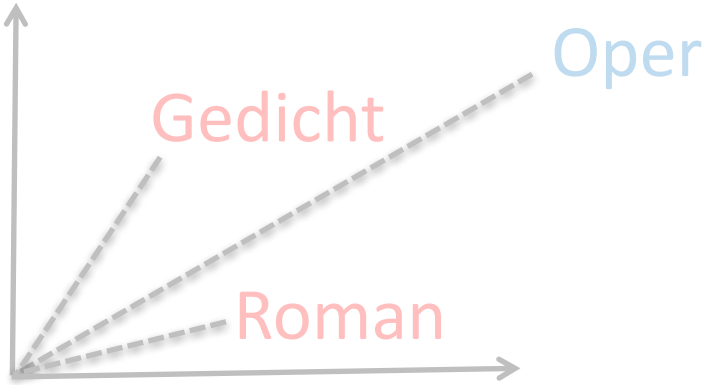
Word Embeddings

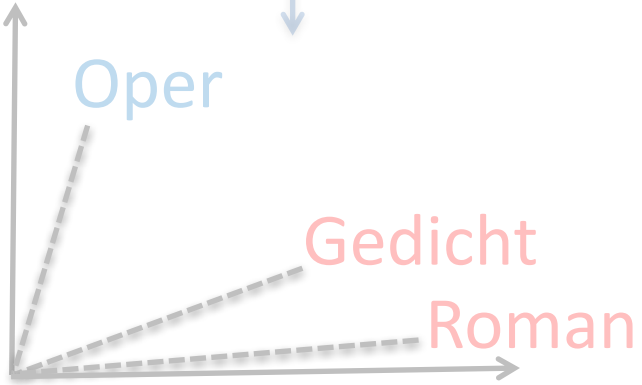
Singulärwertszerlegung

	lesen	Seiten	hören
Gedicht	475	156	76
Roman	823	492	11
Oper	51	19	993



word2vec

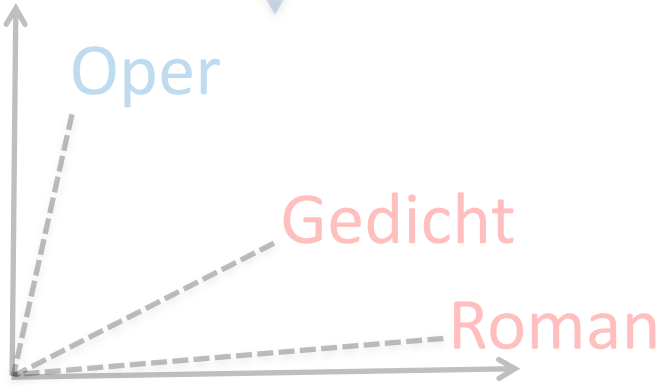




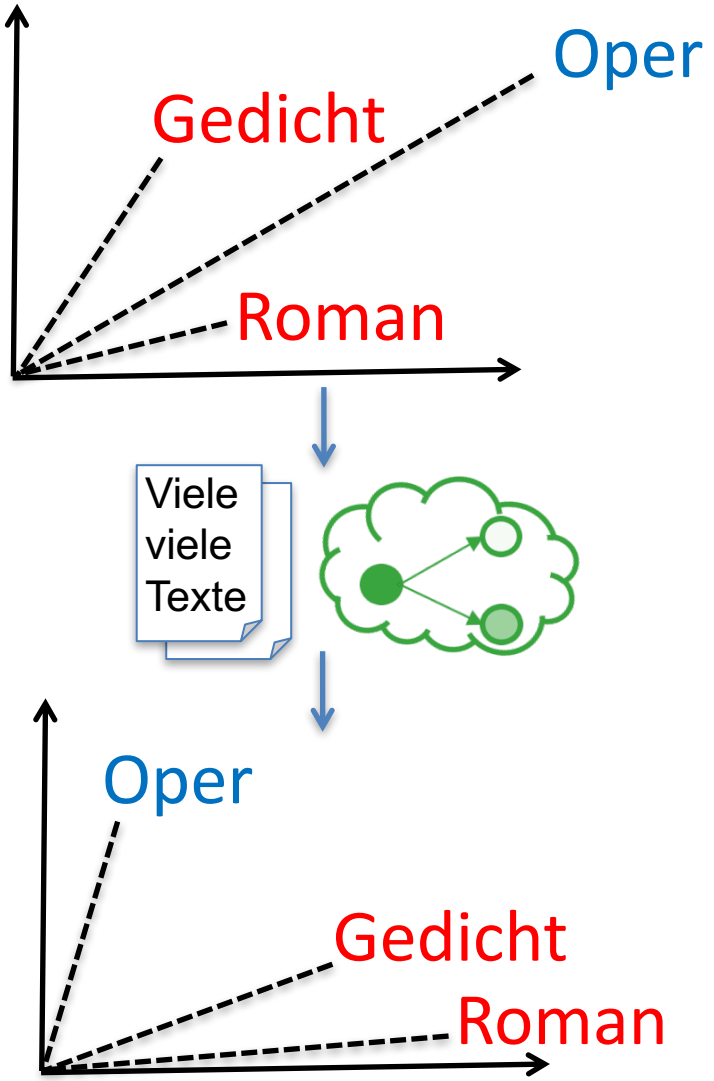
Word Embeddings

Singulärwertszerlegung

	lesen	Seiten	hören
Gedicht	475	156	76
Roman	823	492	11
Oper	51	19	993



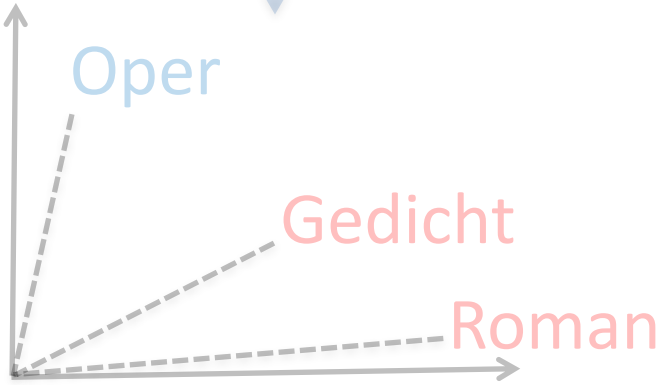
word2vec



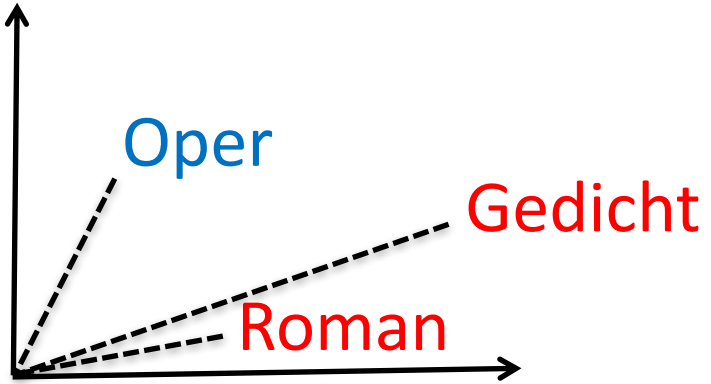
Word Embeddings

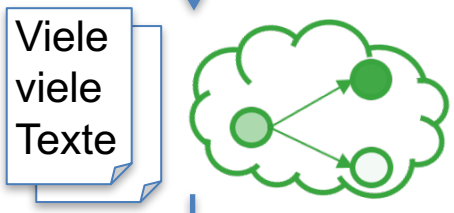
Singulärwertszerlegung

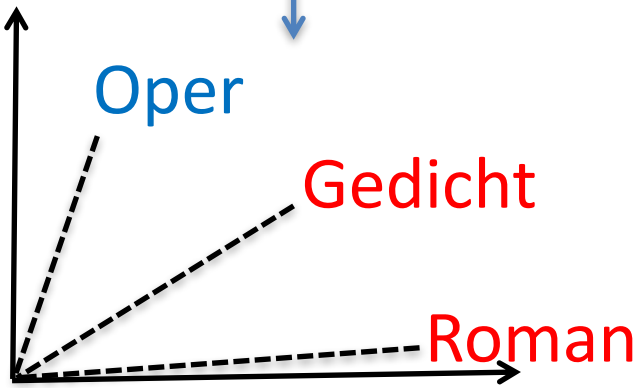
	lesen	Seiten	hören
Gedicht	475	156	76
Roman	823	492	11
Oper	51	19	993



word2vec



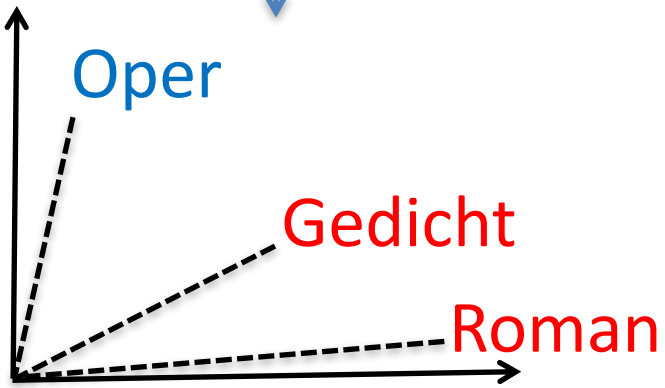




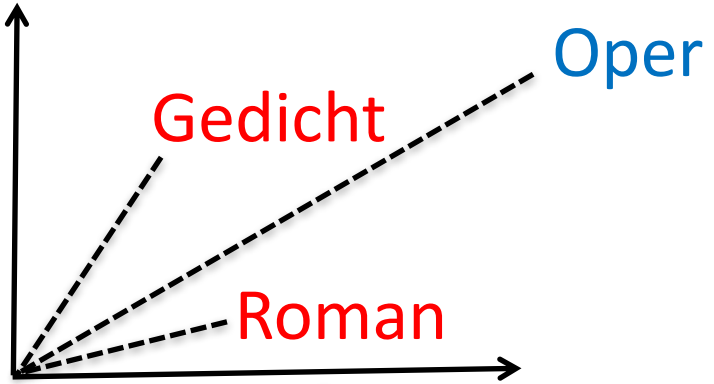
Word Embeddings

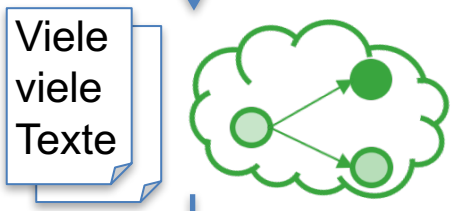
Singulärwertszerlegung

	lesen	Seiten	hören
Gedicht	475	156	76
Roman	823	492	11
Oper	51	19	993



word2vec

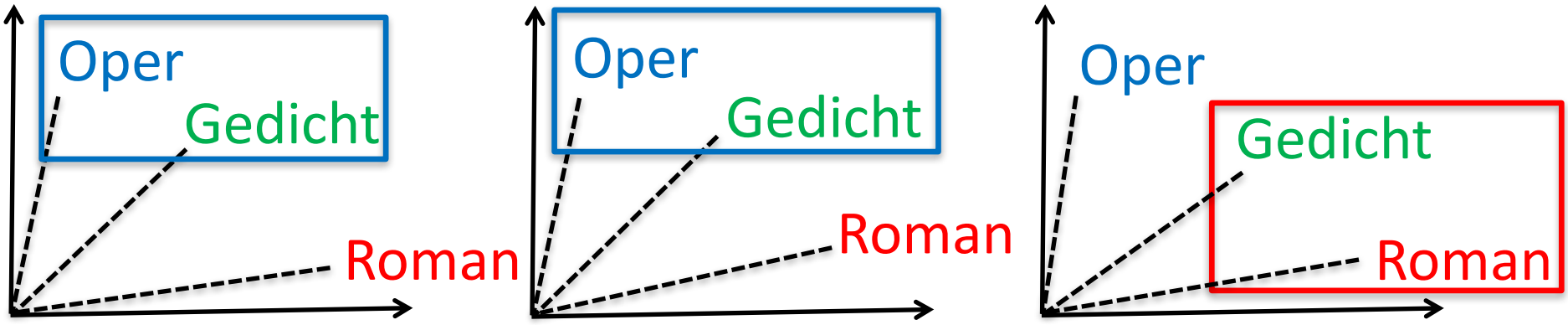




?

Messung der Zuverlässigkeit

- Mehrere Modelle auf gleichen Daten trainieren
- Stimmen sie im ähnlichsten Wort zu einem untersuchten Wort überein?
- Zuverlässigkeit = wie oft Einigkeit über ähnlichstes Wort



Johannes Hellrich & Udo Hahn: Bad Company - Neighborhoods in Neural Embedding Spaces Considered Harmful. In: *COLING 2016*. Osaka, Japan, December 13-16, 2016, pp. 2785-2796.

Ähnlichstes Wort zu „Herz“

Methode	Ähnlichstes	2.-ähnlichstes	3.-ähnlichstes	4.-ähnlichstes	5.-ähnlichstes
word2vec 1	schmerzen	beklommen	busen	bluten	herzen
word2vec 2	bluten	klopfend	busen	beklommen	herzen
word2vec 3	herzen	busen	klopfend	beklommen	bluten
Singulärwerts-zerlegung 1–3	busen	fühlen	liebe	schmerzen	menschen-herz

Johannes Hellrich & Udo Hahn: Don't Get Fooled by Word Embeddings-Better Watch their Neighborhood. In: *Digital Humanities 2017*. Montreal, Canada, August 8-11, 2017. pp. 250-252.

Ähnlichstes Wort zu „Herz“

Methode	Ähnlichstes	2.-ähnlichstes	3.-ähnlichstes	4.-ähnlichstes	5.-ähnlichstes
word2vec 1	schmerzen	beklommen	busen	bluten	herzen
word2vec 2	bluten	klopfend	busen	beklommen	herzen
word2vec 3	herzen	busen	klopfend	beklommen	bluten
Singulärwerts-zerlegung 1–3	busen	fühlen	liebe	schmerzen	menschen-herz

Johannes Hellrich & Udo Hahn: Don't Get Fooled by Word Embeddings-Better Watch their Neighborhood. In: *Digital Humanities 2017*. Montreal, Canada, August 8-11, 2017. pp. 250-252.

Ähnlichstes Wort zu „Herz“

Methode	Ähnlichstes	2.-ähnlichstes	3.-ähnlichstes	4.-ähnlichstes	5.-ähnlichstes
word2vec 1	schmerzen	beklommen	busen	bluten	herzen
word2vec 2	bluten	klopfend	busen	beklommen	herzen
word2vec 3	herzen	busen	klopfend	beklommen	bluten
Singulärwerts-zerlegung 1–3	busen	fühlen	liebe	schmerzen	menschen-herz

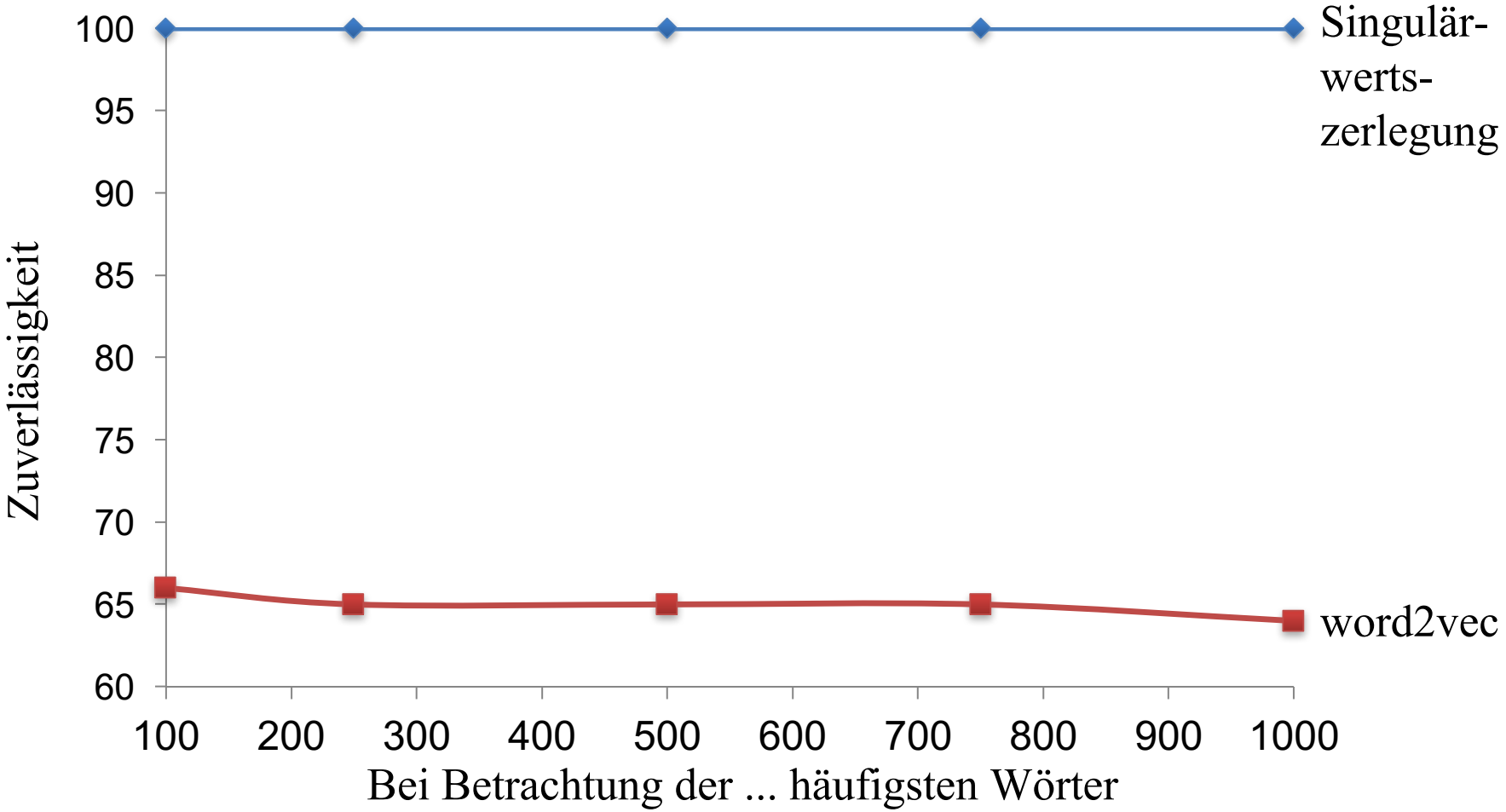
Johannes Hellrich & Udo Hahn: Don't Get Fooled by Word Embeddings-Better Watch their Neighborhood. In: *Digital Humanities 2017*. Montreal, Canada, August 8-11, 2017. pp. 250-252.

Ähnlichstes Wort zu „Herz“

Methode	Ähnlichstes	2.-ähnlichstes	3.-ähnlichstes	4.-ähnlichstes	5.-ähnlichstes
word2vec 1	schmerzen	beklommen	busen	bluten	herzen
word2vec 2	bluten	klopfend	busen	beklommen	herzen
word2vec 3	herzen	busen	klopfend	beklommen	bluten
Singulärwerts-zerlegung 1–3	busen	fühlen	liebe	schmerzen	menschen-herz

Johannes Hellrich & Udo Hahn: Don't Get Fooled by Word Embeddings-Better Watch their Neighborhood. In: *Digital Humanities 2017*. Montreal, Canada, August 8-11, 2017. pp. 250-252.

Zuverlässigkeit und Worthäufigkeit



Fazit

- Die Ähnlichkeit von Wörtern kann durch die Ähnlichkeit ihres Umfelds approximiert werden
- Dadurch kann Wortwandel automatisch nachvollzogen werden
- Methoden müssen auf ihre Tauglichkeit in den Digitalen Geisteswissenschaften hin geprüft werden

Jena University Language & Information Engineering Lab

MODELL ROMANTiK
Variation • Reichweite • Aktualität

Automatische Analyse diachroner Wortsemantik

Johannes Hellrich, M.A.

Digital Humanities in Jena 23.11.2017

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

DFG